DOCUMENT RESUME

ED 190 137 IR 008 655

AUTHOR Penniman, W. D.: Dominick, W. D.
TITLE The Application of Monitor Data Analysis Techniques to Audit Trail Data from an On-Line System.
INSTITUTION Ohio Coll. Library Center, Columbus.
PUB DATE 20 Mar 80
NOTE 19p.: Paper presented at the Mid-Year Meeting of the American Society for Information Science (Pittsburgh, PA, May 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Analysis: Information Retrieval: Information Systems: *Interaction: *Library Networks: *Methods: *Online Systems: Use Studies
IDENTIFIERS *OCLC

ABSTRACT

Summarizing the methods currently being employed at the Ohio College Library Center (OCLC) for user interaction analysis, this study presents suggestions and provides examples for the use of audit trail data available from most online interactive systems. The sample data used in the analysis are from the OCLC online system used for cataloging, interlibrary loans, and other library-oriented operations. The techniques described, however, can be applied to any system which stores that data in an archive file. Tables and references are included. (FM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190137

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Application of Monitor Data Analysis Techniques to Audit Trail Data From an On-line System

by

W. D. Penniman
Director, Development Division, OCLC, Inc.

and

W. D. Dominick
Vice-President, Information Management, Inc.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. David Penniman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

March 20, 1980
OCLC, Inc.
1125 Kinnear Road
Columbus, Ohio 43212

IR008655

63

THE APPLICATION OF MONITOR DATA ANALYSIS TECHNIQUES TO AUDIT TRAIL DATA FROM AN ON-LINE SYSTEM

by

W. D. Penniman
Director, Development Division, OCLC, Inc.

and

W. D. Dominick
Vice-President, Information Management, Inc.

Abstract

Most on-line interactive systems capture an audit trail of user transactions, with associated time stamp and partial user input/system response information, for use in system recovery and accounting. This paper describes techniques which can be applied for further analysis of audit trail data to understand how users are interacting with a particular on-line system.

The OCLC Cataloging Subsystem is used as an example to show how the techniques can be employed. The resulting analysis can lead to improved error diagnosis, system design, and user training as well as a better understanding of the significant variables in effective user interaction with an on-line system.

THE APPLICATION OF MONITOR DATA ANALYSIS TECHNIQUES TO AUDIT TRAIL DATA FROM AN ON-LINE SYSTEM

by

W. D. Penniman
Director, Development Division, OCLC, Inc.

and

W. D. Dominick
Vice-President, Information Management, Inc.

Background

The authors have stressed in previous papers (1,2) the desirability of incorporating an automatic interaction monitor within an on-line information system from the initial design stage. Such a monitor would collect selected data on individual interactions and/or session-level summaries that allow the system designer or data base administrator to evaluate and tune the performance of the system. In most systems today, however, such monitors are not in place and consequently the optimum approach to performance measurement and evaluation is not readily available.

As an alternative to the use of tailored monitor facilities and data, in this paper the authors present suggestions and show examples for the use of audit trail data available from most on-line interactive systems. The sample data used in the analysis presented here is from the OCLC On-line System used for cataloging, interlibrary loan, and other library-oriented operations (3). The techniques, however, can be applied to any system which captures some data on each transaction, and with a time and terminal stamp, stores that information in an archive file. In the case of the OCLC system, data were available from the files which are built daily, include all on-line transactions, and are stored on tape for subsequent use by the Finance Division and more recently by the Research Department (4). This data is now providing valuable insight into how the system users are interacting with various system components, where they are making significant errors, and which portions of the system may require modification or which parts of the user training activities need to be emphasized.

This paper summarizes the techniques currently being employed at OCLC for user interaction analysis, pending the incorporation of an advanced monitor system (5). Sample data illustrate analysis and presentation techniques which should be readily available to all who wish to learn more about how their systems are being used.

Objectives

Given that on-line systems (including OCLC) typically capture "audit trail" data for recovery, security or other purposes, the question for a serious researcher or system designer must be "How can this data be used to understand and/or improve system usage?" The audit trail log is a rich source of data (often too rich) and may actually swamp the analyst with data. In the case of OCLC, the daily file (HISTSTAT - for HISTorical STATistics) contains over one million transactions and often requires more than one 9-track 1600 bpi EBCDIC tape for storage. The Research Department staff was using samples from this vast file to analyze special areas of interest such as the use of particular types of search keys (4,6), but wanted to use the full information content of the transaction records to look at the overall patterns of system usage. This pattern analysis approach is similar to that presented by Penniman (7) for the BASIS on-line retrieval system. Because the vast majority of the transaction data for the OCLC system is associated with on-line cataloging at this time, the analysis focused on this subsystm. The tools developed in the course of this project, however, are generalized and can be applied to other OCLC subsystems (e.g., interlibrary loan) as the need arises.

The basic objectives of the project reported here were to develop the tools and techniques for exploiting the richness of the OCLC audit trail data available through the HISTSTAT files and to use those tools and techniques in subsequent measurement and evaluation of current OCLC system usage. At the detailed level it was essential that the analysis tools be capable of reducing large amounts of data and yet retain the "flavor" of each interactive session. Resolving these two, often conflicting requirements was one of the major challenges of this project.

Methodology

In order to evaluate the processes being invoked by various users and institutions with respect to the OCLC system, a number of concepts need to be defined from an operational viewpoint. The HISTSTAT data collection facility provides enough detailed data that each transaction can be uniquely identified, where a transaction is defined as a user input message to the system and the related system response identifier code. A time-ordered sequence of these transactions represents the process being invoked by the user. Within the OCLC Cataloging Subsystem, it is possible to have an extremely large number of message types (over 40), response codes (up to 17), and eight unique internal system states. This

combined detail can easily obscure what the user is trying to accomplish. By mapping a sequence of messages into a higher-level activity code the process being invoked can be evaluated at a reasonable but not burdensome level of detail. A sequence of messages which corresponds to a continuous interval of time and activity by the user at the terminal is called a session. Each session can be characterized by the type of sequence of acitvities which occurred during that time interval.

The methods incorporated in this preliminary study were meant to be exploratory in nature. Refinements are expected as more data are processed and more subsystems are evaluated. For this study, only the transactions relating to the Cataloging Subsystem were selected for evaluation. This is a simple screening to accomplish because each transaction record has a system number associated with it representing Cataloging, Interlibrary Loan, Serials Control, or any other subsystem being used. By focusing on the most widely used subsystem first, the major problems of data handling would be faced and resolved early on.

Activity types for the Cataloging Subsystem were developed and are presented in Table 1. It is important to note that these groupings can be modified as analysis progresses should a particular set of transactions require greater resolution. The five activities identified in Table 1 exclusively and exhaustively classify all cataloging transactions.

A session has been defined for this preliminary analysis as a sequence of messages issued from a single terminal and bounded by a log-on/log-off message or a ten-minute interval of no activity. The selection of ten minutes as the limit for no activity was arbitrary and can be modified as the timing of interactions becomes more fully understood.

Given an individual session consisting of a start time, end time, and sequence of activities (mapped from transactions) each with a timestamp, the entire pattern and timing of that session can be depicted. This is accomplished via a set of transition matrices, one for transition probabilities between activities and one for elapsed times between activities. This technique has been reported previously by the authors with respect to analysis of other systems (1). The transition analysis can be based on activity couplets (first-order analysis), triplets (second-order) and so forth. Obviously, as the order is increased, so is the complexity and the number of potential unique combinations which must be evaluated. A zero-order analysis would evaluate individual activities without regard to their sequence (e.g., the number of activities of type 1 occurring in a particular session or group of sessions).

The analysis of the temporal aspects of the session can be as fundamental as total elapsed time for each session or

Table 1: Activity Types Developed for
The Analysis of Cataloging Subsystem Transactions

Activity Number	Name	Description
1	Search	All data base searches whether author, title, author/title or identifier number such as ISBN.
2	Select	Selection of an item from a list by entering the number of the item in the list.
3	Record Display	All activities associated with the display of a complete bibliographic record, including calling for a workform or recalling an item from a save area.
4	Produce	Production of cards from an edited record as well as updating or saving a record and producing a spine label.
5	Other	A catch-all category for log-on, end, cancel and other miscellaneous transactions including the display of holdings codes for a given bibliographic record.

as detailed as elapsed time between each activity. The major limitations to detailed timing analysis with the data available are twofold:

1) The clock used to timestamp the transactions from which the activities are derived has a two-second resolution. Consequently the accuracy of the time analysis is limited by this restriction.

2) The timestamp is applied not at the time the transaction is first received by the system nor at the time it is finally shipped to the user, but instead at the time it is released from a queue and routed to a communication processor for final output to the user.

Given these two restrictions it is best not to view the time analysis from a detailed perspective since the elapsed times include not only response time, but also think time, typing time and queuing time which all can vary. The major value of the time analysis is to evaluate the pacing of interactions and to look at the variation (standard deviation and distribution) of this pacing data. The temporal data can be evaluated readily from a first-order perspective (elapsed time between activities) or a zero-order perspective (time spent during a session on a particular type of activity).

The methods developed for this study are similar to those previously tested by Penniman (7, 8, 9) and provide a means for reducing large amounts of data and for comparing individual sessions or groups of sessions to base-line data. The analysis/results presented in the next section of this paper represent initial procedures which will be expanded in the future. It is the authors' hope that the methodology will also be applied to other systems where audit trail data are available.

Data Analysis/Results

A preliminary sample of data was selected from one of the heavy traffic days (February 21, 1979), containing approximately 1,140,000 transactions. These transactions were sorted by terminal identifier and by time. The transactions as stored in HISTSTAT are generally sequenced by time regardless of originating terminal. Because of queuing delays, transactions in the HISTSTAT file may not always be sequenced by time stamp although the vast majority of records are. These occasional anomalies made it necessary to perform sorts on both the time stamp field and the terminal identifier field.

Subsequent to the sorting procedure, the transactions were parsed to determine message type and were screened for subsystems not being evaluated (in this case all

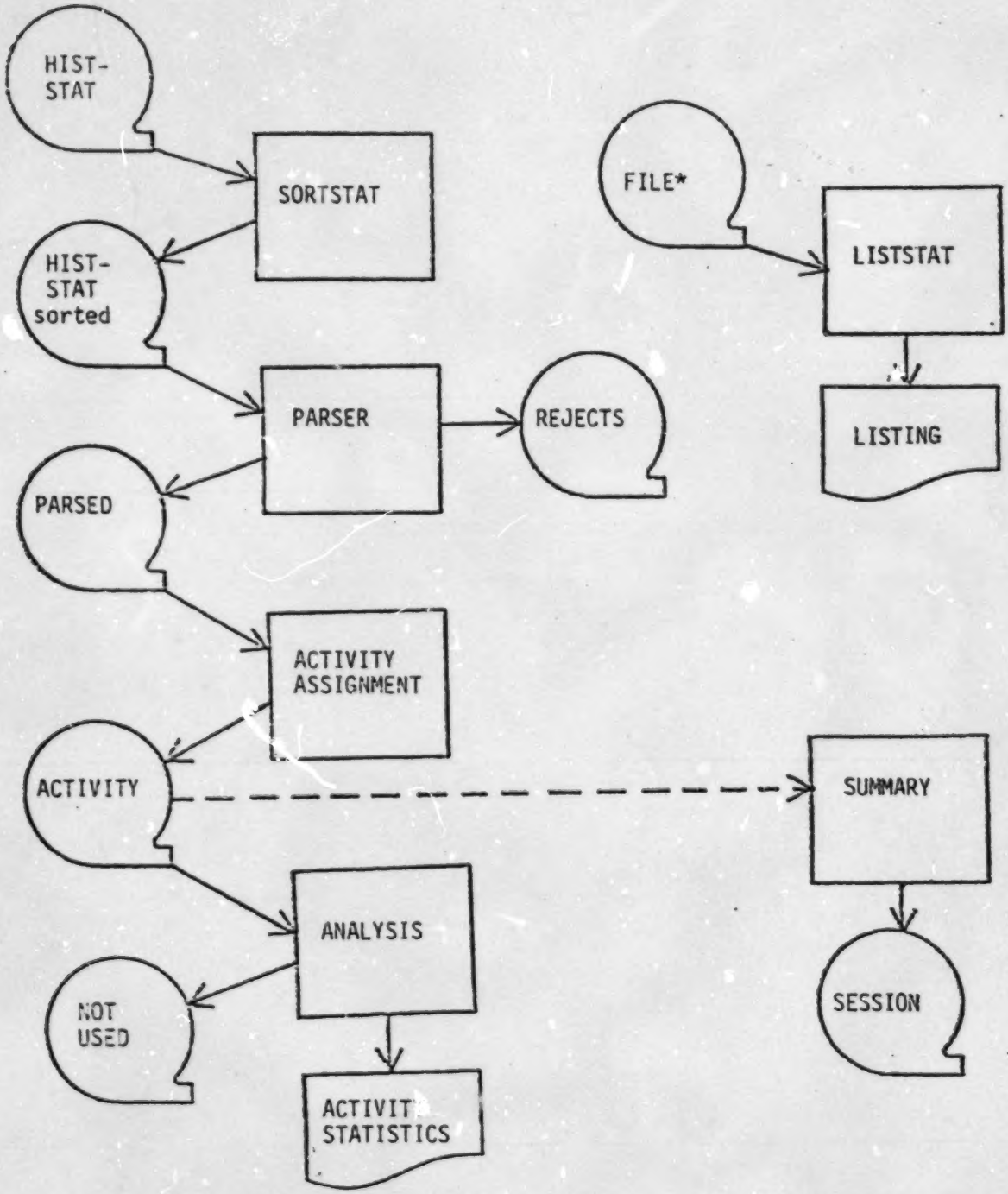
noncataloging transactions). In addition, transactions which resulted in error messages were tagged for special analysis. The selected transition sequences then were processed for activity assignment. Up through this stage, the files used can be retained and listed at each step including the basic HISTSTAT data, the HISTSTAT sorted data, the HISTSTAT sorted data with the parsed information appended (and screened for subsystems of interest) and finally the same records with the activity codes appended. At this point it is possible to perform a detailed transition analysis of the activity sequences or to evaluate aggregate summary data for each session. Figure 1 presents the logical flow of the data analysis steps just described. The types of data analyses and their form of presentation are described in the following paragraphs.

Given a set of transaction records (with associated activity codes) that meet the requirements for a "session" as defined earlier, several analyses can be performed. First, it may be useful to summarize for an entire sample set of sessions the transition probabilities between each of the activity types. This analysis is accomplished by counting the frequency of each transition pair (in the case of first-order analysis) for each session and summarizing these frequency counts across the sessions in the sample. An example of this is shown in Table 2. Each row in Table 2 represents the activity the user has performed in the immediately previous transaction; the columns represent the activity the user performs next. The entries in the TOTAL column represent all transitions from the activity in the corresponding row and are approximately equal to all activities of the corresponding type in the sample. This is approximate because the last activity in each session is not counted as a starting activity.

Probabilities for each couplet are computed by using the frequencies appearing in each cell in Table 2 in conjunction with the respective row total. These transition probabilities for the data in Table 2 are presented in Table 3. Values have been rounded to the nearest one hundredth and rounding errors are indicated by (+) or (-). Note that these values are presented as examples of the data analysis technique (based on a 17,000-transaction sample) and are not necessarily representative of OCLC system usage. Considerably more data need to be analyzed before conclusions can be drawn regarding "typical" OCLC system usage.

The TOTAL frequency counts from Table 2 also can be used to summarize the activities within the sample sessions. For example, Figure 2 illustrates the proportion of each activity type in the sample sessions and is based directly on the frequencies from Table 2. (Note the approximation of total activity frequency due to use of transition totals and not activity totals.) A more accurate analysis can be derived from the session summary records which are described later in this report.

Figure 1: Steps in Processing of HISTSTAT Data for Evaluation of System Usage



*FILE may be any of HISTSTAT, PARSED, ACTIVITY, or REJECTS.

Table 2: Transition Frequency Summary

FREQUENCY OF ACT(I) -> ACT(J) TRANSITIONS

		NEXT ACTIVITY					
		1	2	3	4	5	TOTAL
ACTIVITY	1	2227	1204	1091	112	116	4750
	2	706	1496	319	19	124	2664
	3	54	24	5177	1665	40	6960
	4	1347	3	377	63	36	1826
	5	452	0	73	0	22	547
							16747

Table 3: Transition Probability Summary

PROBABILITY OF ACT(I) -> ACT(J) TRANSITIONS FOR ACT(I)

		NEXT ACTIVITY				
		1	2	3	4	5
ACTIVITY	1	.47	.25	.23	.02(+)	.02(+)
	2	.26	.56	.12	.01	.05
	3	.01	.01(-)	.74	.24	.01(-)
	4	.74	.01	.20	.03	.02
	5	.83	.0	.13	.0	.04

In conjunction with the transition frequency analysis, a companion set of transition matrices are produced which summarize the total interarrival time between each activity type (Table 4) and mean and standard deviation of interarrival time (Table 5) for each transition pair. The total times in Table 4 are presented in hours, minutes and seconds and are summarized for all transitions in the sample data set. The mean and standard deviations in Table 5 are presented in minutes and seconds and are computed from the data in Tables 2 and 4. In addition to this data a graph of time distributions also can be generated for each transition pair. Figure 3 is a sample of this data presentation technique.

A companion analysis to the frequency proportions presented in Figure 2 consists of an analysis of the proportion of time in a session spent in each activity. This data can be derived from the information contained in Table 4. By summarizing the total elapsed time in the sample for each activity transition pair and then computing the proportion by activity type the results presented in Figure 4 are derived. Note that these proportions are slightly different from those presented for the frequency data.

A final data presentation format generated for the sampled sessions provides the following information:

session number

terminal identification number

start time (wall clock)

end time (wall clock)

session elapsed time (hours, minutes, seconds)

total number of messages

message counts for each activity (1 through 5)

A sample of this data is shown in Table 6. The session numbers are assigned arbitrarily to the sessions in this specific sample. The terminal ID is not shown in the sample, but can be used to identify institutional source and location. This summary record provides enough data to perform fundamental (zero-order) data analysis on a large number of sessions and can be used to derive interaction rate analysis and session length distribution analysis (9). Such analysis is planned for a subsequent study, involving a significantly larger sample of data.

It should also be noted that the transition frequency matrices such as shown in Table 2, can be used to statistically test similarities between two or more data sets (8).

Figure 2: Proportion of Activity by Type in Sample Sessions (Based on Frequency Data)

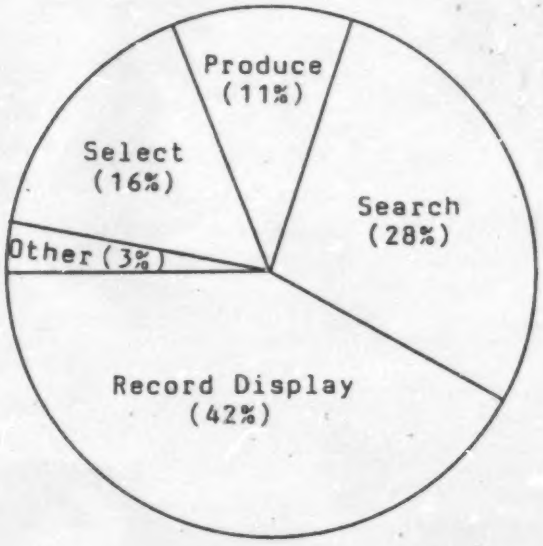


Figure 3: Distribution of Interarrival Times for Activity 1 to Activity 1 Transitions (Partial Sample)

ELAPSED TIME RANGE (minutes:seconds)		FREQUENCY
:00	:10	126
:10	:20	670
:20	:30	463
:30	:40	234
:40	:50	145
:50	1:00	112
1:00	1:10	97
1:10	1:20	85
1:20	1:30	45
1:30	1:40	34
1:40	1:50	30
1:50	2:00	23
2:00	2:10	20
2:10	2:20	22
2:20	2:30	16
2:30	2:40	10
2:40	2:50	11
2:50	3:00	9

Table 4: Total Elapsed Times Between Activities

TOTAL TIME IN ACT(I) BEFORE ACT(I) -> ACT(J) TRANSITION
HOURS : MINUTES: SECONDS

ACTIVITY	NEXT ACTIVITY 1	2	3	4	5
1	27:55:00	9:19:54	12:26:58	2:00:12	1:46:26
2	13:28:16	10:50:08	4:52:22	0:28:24	1:53:56
3	0:38:26	0:07:14	44:48:22	14:57:02	0:50:46
4	12:39:26	0:01:00	2:04:46	0:10:48	0:34:58
5	7:31:36	0:00:00	1:03:20	0: 0:00	0:15:42

Table 5: Mean Elapsed Times (and Standard Deviations) for Activities

MEAN AND STANDARD DEVIATION FOR TIME IN ACT(I)
BEFORE ACT(I) -> ACT(J) TRANSITION
MINUTES : SECONDS
MEAN/(STD DEV)

ACTIVITY	NEXT ACTIVITY 1	2	3	4	5
1	0:45	0:27	0:41	1:04	0:55
	(0:58)	(0:37)	(0:49)	(1:12)	(1:28)
2	1:08	0:26	0:54	1:29	0:55
	(1:07)	(0:34)	(1:15)	(2:03)	(1:06)
3	0:42	0:18	0:31	0:32	1:16
	(0:38)	(0:14)	(0:40)	(0:55)	(1:39)
4	0:33	0:20	0:19	0:10	0:58
	(0:51)	(0:06)	(0:34)	(0:07)	(1:29)
5	0:59	0:00	0:52	0:00	0:42
	(1:26)	(0:00)	(1:33)	(0:00)	(1:09)

Figure 4: Proportion of Time Spent in Each Activity Type in Sample Sessions

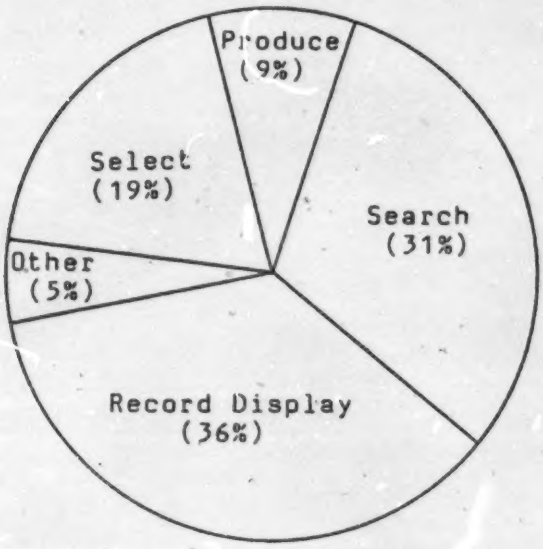


Table 6: Example of Session Summary Records

```
SESSION # = 533  TERMINAL ID = XXXX  START TIME = 8:26:58  END TIME = 8:42:22  SESSION TIME = 0:15:24
TOTAL MSGS = 16  ACT(1) MSGS = 10  ACT(2) MSGS = 0  ACT(3) MSGS = 1  ACT(4) MSGS = 0  ACT(5) MSGS = 5

SESSION # = 534  TERMINAL ID = XXXX  START TIME = 8:42:48  END TIME = 8:45:56  SESSION TIME = 0:03:08
TOTAL MSGS = 4  ACT(1) MSGS = 1  ACT(2) MSGS = 0  ACT(3) MSGS = 2  ACT(4) MSGS = 0  ACT(5) MSGS = 1

SESSION # = 535  TERMINAL ID = XXXX  START TIME = 8:46:24  END TIME = 8:46:46  SESSION TIME = 0:00:22
TOTAL MSGS = 4  ACT(1) MSGS = 1  ACT(2) MSGS = 0  ACT(3) MSGS 2  ACT(4) MSGS = 0  ACT(5) MSGS = 1

SESSION # = 536  TERMINAL ID = XXXX  START TIME = 8:47:08  END TIME = 8:47:14  SESSION TIME = 0:00:06
TOTAL MSGS = 2  ACT(1) MSGS = 1  ACT(2) MSGS = 0  ACT(3) MSGS = 1  ACT(4) MSGS = 0  ACT(5) MSGS = 0

SESSION # = 537  TERMINAL ID = XXXX  START TIME = 8:48:00  END TIME = 8:48:00  SESSION TIME = 0:00:00
TOTAL MSGS = 1  ACT(1) MSGS = 0  ACT(2) MSGS = 0  ACT(3) MSGS = 1  ACT(4) MSGS = 0  ACT(5) MSGS = 0

SESSION # = 538  TERMINAL ID = XXXX  START TIME = 8:39:56  END TIME = 8:44:42  SESSION TIME = 0:04:46
TOTAL MSGS = 8  ACT(1) MSGS = 0  ACT(2) MSGS = 0  ACT(3) MSGS = 7  ACT(4) MSGS = 1  ACT(5) MSGS = 0
```

This ability to quantitatively compare interaction patterns or processes is a significant step towards evaluating and understanding the critical variables in effective on-line user interaction.

Conclusion and Recommendations

The data analysis techniques presented in this paper provide an effective means of reducing the large quantity of data collected through audit trail logs. The techniques offer the flexibility necessary to evaluate interaction data across a wide range of system applications using basically identical software. The current software, written in PASCAL and running on a Xerox Sigma 9 computer under the CP-V operating system, will be used to analyze historical statistics data for all OCLC subsystems. The activity assignment algorithms will be modified in the activity assignment software and transactions in the subsystem of interest rather than just the Cataloging Subsystem will be selected and parsed.

Further analyses of cataloging data will include a significantly larger sample (several million transactions) and will include comparisons within and across terminals. These analyses will focus on errors and error patterns as well as implicit requirements for user training. Preliminary indications from even the small amount of data presented here show that many transactions that result in error responses from the system are correct in form, but terminated with an improper termination key. This pattern will be analyzed more fully in subsequent research.

Finally, it is important to note that analysis of strings of activities (i.e., beyond the two-step process depicted here) will provide a much better understanding of typical and atypical sessions within the OCLC environment.

Acknowledgments

This project was supported entirely by OCLC, Inc., within the recently formed OCLC Research Department. The authors are indebted to Marvin Watkins and Keith Johnston, Research Assistants within the Research Department responsible for the data analysis software preparation and software documentation respectively. Their efforts in coding, testing, and describing the software used for the analyses depicted within this paper were outstanding. Dr. Dominick serves as consultant to the Research Department and has provided significant contributions in the human/computer interaction research program which includes the on-line monitoring research project currently underway.

References

1. Penniman, W. D. and W. D. Dominick, "Monitoring and Evaluation of On-Line Information System Usage," accepted for publication in Information Processing and Management, (to appear).

2. Dominick W. D. and W. D. Penniman, "Automated Monitoring to Support the Analysis and Evaluation of Information Systems," Proceedings of the Second International Conference on Information Storage and Retrieval, ACM SIGIR FORUM, Vol. XIV, No. 2, September 27-28, 1979, pp.2-9.

3. Allison, A. M. and A. M. Allan, (editors), OCLC: A National Library Network, Enslow Publisher, 1979.

4. Rastogi, K. B., "Maximum and Minimum Derived Search Key Use and Its Relationship With Found and Not-Found System Responses," Journal of Library Automation, Vol. 12, June, 1978, pp. 169-175.

5. Dominick, W. D. and W. D. Penniman, "Interaction Monitoring Considerations Within Network-Based Information Systems," North American Networking Collected Papers, Eighth Mid-Year Meeting of the American Society for Information Science, Banff, Alberta, Canada, May 16-19, 1979, p. F4.

6. Rastogi, K. B., "Retrieval Behavior of Derived Truncated Search Keys for a Large On-Line Bibliographic File," accepted for publi cation in the Journal of the American Society for Information Science, (to appear).

7. Penniman, W. D., Rhythms of Dialogue in Human-Computer Conversation, Unpublished PhD Dissertation, The Ohio State University, Columbus, OH, 1975.

8. Penniman, W. D., "A Stochastic Process Analysis of On-Line User Behavior," Proceedings of the American Society for Information Science, Vol. 12, 1975, pp. 147-148.

9. Penniman, W. D. and J. C. Perry, "Tempo of On-Line User Interaction," Compendium of Papers, Fifth Mid-Year Meeting of the American Society for Information Science, May, 1976, pp. 10-23.